科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１６　日現在

機関番号：　１３９０４
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００１６７

研究課題名（和文）半教師付き距離学習によるオンラインデータ分類

研究課題名（英文）Online Data Classification by Semi-Supervised Metric Learning

研究代表者
岡部　正幸（OKabe, Masayuki）

豊橋技術科学大学・情報メディア基盤センター・助教

研究者番号：５０３６２３３０

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：距離学習は，ネット上のニュース記事のカテゴリ分けやショッピングサイトにおける商品の推薦など様々なデータを機械的に分類するアルゴリズムの構築に役立つ要素技術である．本研究では，この距離学習における問題点として，訓練データ獲得のための人的コストと学習に要する計算コストを取り上げ，これらを削減するための各種アルゴリズムを構築した．また，応用例として，外れ値検出システムによるネットワークトラフィックデータからの異常発見システムを構築した．

研究成果の概要（英文）：Metric learning is a kind of method useful for automatic data classification such as categorization of news on the Web or item recommendation in online shopping. In this research, we deal with two problems of costs for acquiring training data and computation when using metric learning, and propose some algorithms to resolve the problems. We also construct an anomaly detection system from network traffic data based on the proposed algorithms.

研究分野：　総合領域

科研費の分科・細目：　情報学・知能情報学

キーワード：　知能情報処理　機械学習　データマイニング

１．研究開始当初の背景

機械学習において，データ間の距離（または類似度）をどのように計算するかは，データ分類・クラスタリングの精度に直結する極めて重要な要素である．距離学習は，予め類似性の判定を行ったデータペアを制約知識として利用することにより，データ集合全体の距離尺度を恣意的に変化させ，分類精度向上を図る技術である．距離学習は，広くは特徴ベクトルにおける次元削減，属性選択技術の一種と捉えることができるが，主成分分析や判別分析などの従来手法に比べ分類精度向上における性能が高いことが近年の研究により明らかになっている．また，距離学習はクラス分類・クラスタリングにおける種々の学習アルゴリズムに組み込むことが可能な汎用性の高い要素技術であり，データマイニング・情報検索・画像認識・統計的言語処理など多岐に渡る応用分野での利用が期待されている．

距離学習の性能は一般に学習の際に利用する制約ペア数に比例して良くなる傾向にあるが，データ数に対するペア候補数は２乗のオーダであるので，性能向上には非常に多くの制約ペアが必要となる．このため，データ数が大規模になると制約ペア生成における類似性判定のための人的コストが問題となる．また，これまでの距離学習アルゴリズムの多くは，計算コストが高くデータ更新時の再学習が容易ではない．更に，近年の研究から距離学習の性能は制約ペア集合の量だけでなくその質にも大きく影響を受けることが分かってきており，制約ペアをうまく取捨選択し，限られた制約ペア数で効率的に学習を行う方法が求められている．このように，距離学習を実用的な応用システムに組み込むには，スケール化及び学習効率の観点から解決すべき問題があった．

２．研究の目的

本研究では，距離学習を実用システムで利用する際の２つの懸念，制約ペア判定における人的コストと学習時における計算コストの問題を取り除くため，利用可能な制約ペア集合が少数に限定された環境においても効果の高い距離学習が行える半教師付き学習アルゴリズムの構築とオンライン上で頻繁にデータ更新が行われる環境においても効率的に再学習が行えるための逐次学習アルゴリズムの構築を行う．また，実用システムへの応用展開として，構築したアルゴリズムを組み込んだクラスタリングベースの外れ値検出システムを作成し，ネットワークトラフィックデータからの異常発見問題への適用・評価を行う．

３．研究の方法

まず，半教師付き距離学習アルゴリズムの構築については，制約を持つデータペアの関係性を近傍データへと伝搬させる方法について検討する．また，距離学習をデータペアが類似しているか否かという２値分類問題とみなした場合の方法についても検討する．

次に，逐次更新可能な距離学習アルゴリズムの構築については，まず，距離学習における計算コストを削減する方法として，距離学習における変換行列やカーネル行列の半正定値性の判定を効率的に行える方法について検討する．次に，制約ペア集合から不要なものを除去し，期待効用の高い制約ペア候補を選別することで効率的に学習を行うため，能動学習的なアプローチによる制約ペアの期待効用に基づく選択方法について検討する．

最後に，実用システムへの応用展開については，外れ値検出システムによるネットワークトラフィックデータからの異常発見システムについて検討する．外れ値検出は構築した距離学習アルゴリズムに基づくクラスタリングによって行う．これを学内ネットワークからキャプチャしたトラフィックデータに適用する．

４．研究成果

(1)　グラフカットに基づく制約付きクラスタリング

本研究では，半教師付き距離学習アルゴリズムの一種である半正定値計画に基づく制約付きグラフカットを提案し，これを用いたクラスタリングアルゴリズムを構築した．提案手法では，must リンク制約を用いたグラフカット問題を半正定値計画として定式化し，得られた近似解を用いてクラスタリングを行う．この近似解として得られる行列の要素は各データ対と対応しており，その値は対応するデータ対が同一クラスタに属するかどうかを判定するための一種の距離行列として使用することができる．このため提案手法では，得られた解行列を分解することなく，実際のクラスタ分割を効率的に行うことができる．

実験では，提案手法を他の３つの代表的な従来手法と比較し，UCI レポジトリと CLUTO ベンチマークの半分以上のデータセットにおいて同等またはそれ以上の性能を持つことを示した．また，実行時間についても，データサイズが大きい場合でも同じ SDP を利用した従来手法と比較して最大２倍程度の計算時間で抑えることができることを示した．

(2)　カーネル行列学習に基づくクラスタアンサンブル

本研究では，計算コストの低い制約付きクラスタリングアルゴリズムの一つである COP-Kmeans に着目し，そのクラスタリング性能を補うため，アンサンブル学習の原理を利用したクラスタアンサンブル法を提案する．提案方法は，制約をその優先度順に充足させるよう修正された COP-Kmeans アルゴリズムを弱学習アルゴリズムとして，ブースティン

グの原理により優先度を制御することで，充足された制約集合の異なる複数のクラスタリング結果を統合する．

実験では，提案手法と従来手法のクラスタリング性能と計算時間について 12 種類のデータセットにおいて比較し，提案手法が多くのデータセットにおいて計算時間が少なくかつ同等以上の性能を持っていることを示した．また，提案手法ではクラスタリング性能がブースティングステップ数の増加とともに向上することを示し，ブースティングステップ数を適切に設定することで計算時間を必要最小限に抑えることが可能であることがわかった．

(3) 不確実性サンプリングに基づく能動的制約獲得

本研究では，制約付きクラスタリングの性能を向上させる制約の洗濯方法について，クラスタアンサンブルおよび不確実性サンプリングに基づき能動的に行う方法を提案する．提案手法は，カーネル行列学習によるクラスタアンサンブルをベースとしている．このクラスタアンサンブルでは，制約付き K-means クラスタリングアルゴリズムをデータ対の制約ラベルを判定する弱学習器として利用することにより，制約の重みを制御しながら生成される複数のクラスタリング結果を統合する．提案手法では，クラスタリング結果の向上に有効な制約を選択する基準として，各クラスタリング結果において同一クラスタおよび別クラスタに分類される頻度の不確実性を利用する．

提案手法の有効性を検証するため，UCI レポジトリおよび CLUTO データセットを用いた実験を行ったところ，ランダムな選択方法よりも半分以上のデータセットにおいて有効性を示すことができた．また，制約選択においては，制約を複数個まとめて選択しバッチ的に処理するよりも少数ずつ対話的に選択することが良いということも明らかになった．

(4) 制約選択を支援するインタラクションデザイン

本研究では，ユーザにクラスタリング結果を提示して，ペアワイズ制約を与えてもらい，制約付きクラスタリングを繰り返すインタラクティブ制約付きクラスタリングの枠組みにおいて，人間の制約選択を支援するインタラクションデザインを提案した．具体的な方法として，制約効果を顕在化する GUI とクラスタリング結果を複数視点から俯瞰することができる GUI の機能を導入したシステムを提案し，実装した（図 1）．制約効果の顕在化は，与えられた制約の影響を受けたと考えられるデータを強調表示することでユーザに把握しやすくする．また，複数視点の変更機能は，2 次元表示する 2 軸を多次元尺度構成法の固有値の組み合わせで変化させることで，複数視点を切り替える機能である．

実装したシステムを用いて，画像データのクラスタリングを対象とした参加者実験を行い，提案する GUI のないシステム，計算論的能動学習，能動学習なしシステムとのパフォーマンスの比較および制約選択の戦略に関するアンケート調査を行った．その実験結果から，提案する GUI が人間の制約選択の支援に効果があることを実験的に示した．そして，実験から得られた人間の制約選択の戦略，計算論的能動学習との関係を考察した．

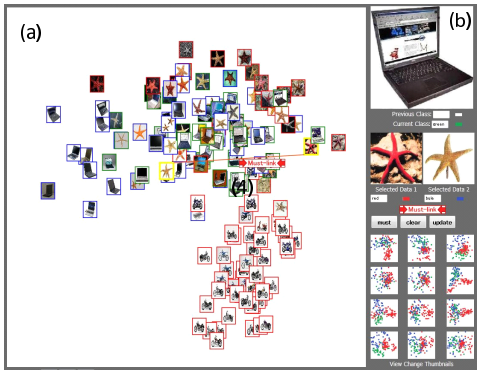

図 1 ユーザインタフェース

(5)

本研究では，外れ値検出に基づくファイアウォールログからの異常検知を効率的に行うためのインタラクション設計について考察し，特徴選択を行うためのユーザインタフェースを持つ検知システムを構築した（図 2）．提案システムは，組織内と組織外の境界に設置されたファイアウォールから出力されるネットワークトラフィックログをソースとして，異常通信の対象となっている組織内のホストを外れ値検出によって特定するための機能を持つ．外れ値検出を効率的に行うには，検知対象ごとに適した特徴量を選択することが重要となるため，本システムでは外れ値ランキングの可視化機能を提供することにより，ユーザはシステムと対話的に特徴選択を行うことができる．

試験運用における定性的評価を行ったところ，本システムによりファイル交換ソフトの使用などが発見できることが分かった．

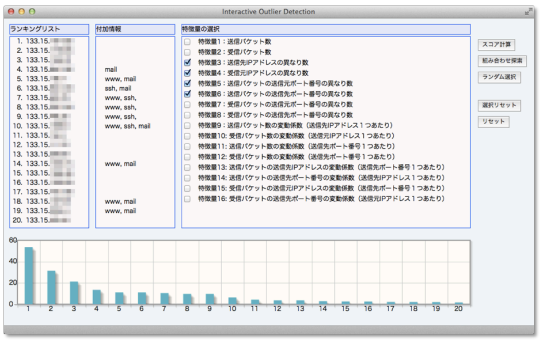

図 2 ユーザインタフェース

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　5　件）

① Masayuki Okabe and Seiji Yamada, “Active Sampling for Constrained Clustering”, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.18, No.2, pp.232-238 (2014) （査読有）
http://www.fujipress.jp/finder/xslt.php?mode=present&inputfile=JACII001800020017.xml

② 山田誠二, 水上淳貴, 岡部正幸 “制約付きクラスタリングにおける人間の能動学習を促進するインタラクションデザイン”, 人工知能学会論文誌, Vol.29, No.2, pp.259-267 (2014) （査読有）
http://dx.doi.org/10.1527/tjsai.29.259

③ Masayuki Okabe and Seiji Yamada, “Graph-cut based Constrained Clustering by Grouping Relational Labels”, Journal of Emerging Technologies in Web Intelligence, Vol.4, No.1, pp.43-50, (2012) （査読有）
DOI: 10.4304/jetwi.4.1.43-50

④ 岡部正幸, 山田誠二, “制約付きグラフカットによる逐次クラスタリング”, 人工知能学会論文誌, Vol.27, No.3, pp.193-203, (2012) （査読有）
http://dx.doi.org/10.1527/tjsai.27.193

⑤ Masayuki Okabe and Seiji Yamada, “An Interactive Tool for Human Active Learning in Constrained Clustering”, Journal of Emerging Technologies in Web Intelligence, Vol.3, No.1, pp.20-27, (2011) （査読有）
DOI:10.4304/jetwi.3.1.20-27

〔学会発表〕（計　6　件）

① 岡部正幸, 山田誠二, “外れ値検出に基づく対話的ファイアウォールログ分析”, 第 28 回人工知能学会全国大会, 3B3-OS-10a-4, 2014/5/14, 愛媛

② Masayuki Okabe and Seiji Yamada, “Uncertainty Sampling for Constrained Cluster Ensemble”, Conference on Technologies and Applications of Artificial Intelligence (TAAI 2013), pp.257-262, 2013/12/6, 台北

③ Masayuki Okabe and Seiji Yamada, “Clustering by Learning Constraints Priorities”, The IEEE International Conference on Data Mining (ICDM2012), pp.1050-1055, 2012/12/13, ブリュッセル

④ Masayuki Okabe and Seiji Yamada, “Clustering with Extended Constraints by Co-Training”, International Workshop on Intelligent Web Interaction, pp.79-82, 2012/12/4, マカオ

⑤ Masayuki Okabe and Seiji Yamada, “Active Sampling for Constrained Clustering”, International Conference on Soft Computing and Intelligent Systems (SCIS&ISIS 2012), pp.399-402, 2012/11/21, 神戸

⑥ Masayuki Okabe and Seiji Yamada, “Graph-cut based Iterative Constrained Clustering”, International Workshop on Intelligent Web Interaction, pp.126-129, 2011/8/22, リヨン

６．研究組織
(1)研究代表者
岡部　正幸（OKABE, Masayuki）
豊橋技術科学大学・情報メディア基盤センター・助教
研究者番号：50362330